

Lecture 7 : 확률

Fastcampus Math Camp

신승우

Saturday 7th July, 2018

- 1 확률의 정의
- 2 확률변수
- 3 조건부확률의 활용
- 4 Different Perspectives on Probability

확률의 정의 I

우리가 어떤 실험에 앞서서, 나올 수 있는 모든 가능한 결과는 알지만 그 각각 중 무엇이 나올지 모른다고 생각하자. 나올 수 있는 모든 가능한 결과의 집합을 **표본공간(Sample Space)**이라 한다. 예컨대,

- 동전을 던짐 : $S_1 = \{Head, Tail\}$
- 주사위를 던짐 : $\{1, 2, 3, 4, 5, 6\}$
- 두 개의 동전을 던짐 : $S_1 \times S_1$
- 0부터 1 사이의 숫자를 아무렇게나 고름 : $\{x | 0 \leq x \leq 1\}$

라고 할 수 있다.

확률의 정의 II

이 때, 표본공간의 부분집합을 **사건(Event)**라고 한다. 예컨대, 집합 S_1 의 부분집합은 총 4개로, $\{\}$, $\{Head\}$, $\{Tail\}$, $\{Head, Tail\}$ 이다. 이 부분집합 각각을 사건이라 하며, 어떤 사건이 일어났다는 것은 관측 결과가 그 사건의 원소임을 의미한다. 예컨대, 동전을 던져서 앞면(Head)가 관측된다면, 두 이벤트 $\{Head\}$, $\{Head, Tail\}$ 이 일어난 것이다. 만약 두 사건 E, F의 교집합이 공집합이라면, 이 때 E, F를 **배반사건 (Mutually Exclusive)**라 한다.

확률의 정의 III

위에서 언급된 용어들을 이용하여 확률을 정의할 수 있다.

확률

확률은 사건을 실수에 대응시키는 함수 P 중 다음 3가지 조건을 만족하는 것을 말한다.

- $\forall E \subset S, 0 \leq P(E) \leq 1$
- $P(S) = 1$
- $M = \{E_i | \forall E_i, E_j \in M, E_i \cup E_j = \phi\}, P(\cup_{e \in M} e) = \sum_{e \in M} P(e)$

예컨대,

- 양수 p 에 대해서 $P(\{Head\}) = p$ 이고 $P(\{Tail\}) = 1 - p$ 라면 P 는 동전던지기 실험에 대한 확률이다.
- 임의의 유한한 표본공간 S 에 대해서, 함수 $P(E) = \frac{|E|}{|S|}$ 는 항상 확률이다.

내일의 날씨를 예측할 때, 전날 밤에 달무리가 관측(이 사건을 M이라 하자)되었으면 다음 날 비가 올(이 사건을 R이라 하자) 확률이 높을 것이다. 예를 들어서 다음과 같이 생각할 수 있다.

- M이 일어날 확률을 0.2라 하자.
- M이 일어남 \rightarrow R이 일어날 확률 0.9
- M이 일어나지 않음 \rightarrow R이 일어날 확률 0.3

조건부확률 II

이 때, M과 R이 일어나거나, 일어나지 않을 경우에 따라서 각각 4 가지 경우를 생각할 수 있다. 이를 표를 이용하여 확률계산을 해 보자.

	M	not M	prob
R	$0.2 \times 0.9 = 0.18$	$0.8 \times 0.3 = 0.24$	$P(R) = 0.42$
not R	$0.2 \times 0.1 = 0.02$	$0.8 \times 0.7 = 0.56$	$P(\text{not } R) = 0.58$
prob	$P(M) = 0.2$	$P(\text{not } M) = 0.8$	1

위와 같이 계산할 수 있다. 이 때, M이 일어났을 때 R이 일어날 확률 (조건부확률)과, M과 R이 같이 일어날 확률을 구분하여 생각하여야 한다. 전자의 조건은 표본집단이 M이 일어난 집단으로 바뀌고, 후자의 경우는 전체 표본집단 하에서의 확률을 생각하기 때문이다. 즉, 2행 2열의 확률은 M과 R이 같이 일어난 확률이지만, 이를 $P(M)$ 으로 나누면 0.9가 되어 원 조건을 만족하게 된다.

조건부확률 III

이에서 다음의 공식을 생각할 수 있다.

조건부확률

F가 일어났다는 선행조건 하에서 F가 일어날 확률을 $P(E|F)$ 라고 쓰며,

$P(E|F) = \frac{P(EF)}{P(F)}$ 로 계산한다.

조건부확률 IV

이 때, 조금 다른 상황을 생각해 보자. 길동이가 노란 옷을 입는 사건을 Y 라고 하면, 이 때 일반적으로 길동이의 옷 색과 비가 내릴 확률은 상관이 없을 것이므로 위 상황과 같은 확률로 비가 온다고 생각하면, 다음이 성립할 것이다.

- R 이 일어날 확률은 0.42, Y 가 일어날 확률은 0.1
- Y 가 일어남 $\rightarrow R$ 이 일어날 확률 0.42
- Y 가 일어나지 않음 $\rightarrow R$ 이 일어날 확률 0.42

조건부확률 V

이 때는 위와 비슷하게 아래 표를 생각할 수 있다.

	Y	not Y	prob
R	$0.1 \cdot 0.42 = 0.042$	$0.9 \cdot 0.42 = 0.24$	$P(R) = 0.42$
not R	$0.9 \cdot 0.58 = 0.522$	$0.1 \cdot 0.58 = 0.058$	$P(\text{not } R) = 0.58$
prob	$P(Y) = 0.1$	$P(\text{not } Y) = 0.9$	1

즉, 이렇게 상관없는 두 사건의 경우, $P(E \cap F) = P(E)$ 이다. 이를 두 사건이 서로 **독립사건**이라고 한다. 이 경우, 위 식에 의해서 $P(EF) = P(E)P(F)$ 이다.

조건부확률 VI

위 두 경우에서 일반화시켜, 사건 E 와 표본공간을 분할하는 F_i 들에 대해서 $P(E|F_i)$ 들이 알려져 있다고 할 때, $P(F_j|E)$ 를 다음과 같이 계산할 수 있다. 이 때 분할은 다음의 두 가지를 의미한다.

- F_i 들의 합집합이 표본공간이 된다.
- F_i 끼리는 교집합이 모두 공집합이다.

이 때, 어떤 표본공간의 부분집합 E 에 대해서 다음이 성립한다.

$$P(E) = \cup P(EF_j) \quad (1)$$

$$P(E) = \sum P(EF_j) \quad (2)$$

$$= \sum P(E|F_j)P(F_j) \quad (3)$$

조건부확률 VII

또한, 어떤 특정한 F_j 에 대해서

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} \quad (4)$$

이므로, 다음이 성립한다.

$$P(F_j|E) = \frac{P(EF_j)}{\sum P(E|F_j)P(F_j)} \quad (5)$$

이를 베이즈의 정리라고 한다.

Outline

- 1 확률의 정의
- 2 확률변수
- 3 조건부확률의 활용
- 4 Different Perspectives on Probability

실험 값 자체가 중요한 경우도 있지만, 가끔은 실험 값을 가공한 결과값이 더 중요할 때가 있다. 예컨대, 주사위 두 개를 던져서 그 합의 짝/홀로 승패를 가르는 게임을 한다고 하면, 주사위 각각의 결과값은 크게 중요하지 않을 것이다.

이런 결과값은 수학적으로 볼 때, 표본공간의 각 원소들에 대한 함수라고 볼 수 있다. 이런 함수를 **확률변수(Random Variable)**이라고 한다.

확률변수

표본공간을 정의역으로, 실수를 공역으로 가지는 **함수**를 그 표본공간에서의 확률변수라고 한다.

확률변수의 예시로, 동전을 앞면이 나올 때까지 던지는 횟수를 생각하자. 그러면, 표본공간 S 와 확률변수 N 은 다음과 같이 정의된다.

- $S = \cup_{i=1}^n F_i$
- $F_i = \{(T, T, \dots, T, H)\}$
- $P(F_i) = (1 - p)^{i-1}p$
- $N(F_i) = i$

이 때, p 는 동전을 던져서 앞면이 나올 확률이다.

이산 확률변수(Discrete Random Variable)

가산집합이거나 유한집합인 표본집합 위에서의 확률변수를 이산 확률변수라고 한다.

이 때, 이산 확률변수 X 에 대해서

- Probability Mass Function(PMF) : $f(a) = P(X = a)$
- Cumulative Distribution Function(CDF) :

$$F(a) = P(X \leq a) = \sum_{x \leq a} f(a)$$

가 정의된다. 이산 확률분포는 보통 PMF를 이용하여 정의된다. 예컨대, 앞장에서의 동전을 던지는 실험을 보면, $f(i) = (1 - p)^{i-1}p$ 가 pmf이다. 이 때, 정의역 구간 내에서 pmf의 합은 1이 되어야 한다. 이는 저 식으로 1 부터 무한대까지 더하는 극한을 통해 알아볼 수 있다. 위와 같이, p 같은 변수를 보통 parameter이라고 부른다.

다양한 이산 확률변수들 I

- 베르누이 확률변수

베르누이 확률변수

pmf f 가 다음과 같이 주어지는 확률변수를 베르누이 확률변수라 한다.

$$f(0) = P(X = 0) = 1 - p \quad (6)$$

$$f(1) = P(X = 1) = p \quad (7)$$

이는 어떤 성공할 확률이 p 인 시행이 성공하면 1, 실패하면 0인 확률변수라고 볼 수 있다. 이것이 n 번 시행되는 경우를 Binomial Random Variable이라고 한다. 이 때의 pmf f 는 시행횟수 n 과 성공확률 p 에 대해서 다음과 같다.

$$f(i) = \binom{n}{i} p^i (1 - p)^{n-i} \quad (8)$$

다양한 이산 확률변수들 II

- 기하학적 확률변수

기하학적 확률변수

기하학적 확률변수는 pmf f 가 parameter p 에 대해서 다음과 같이 주어지는 확률변수를 뜻한다.

$$f(i) = P(X = i) = (1 - p)^{i-1} p \quad (9)$$

- 푸아송 확률변수

푸아송 확률변수

푸아송 확률변수는 pmf f 가 파라미터 λ 에 대해서 다음과 같이 주어지는 확률변수를 뜻한다.

$$f(i) = P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!} \quad (10)$$

연속 확률변수

비가산집합인 표본공간 하에서의 확률변수를 연속 확률변수라고 한다.

연속 확률변수의 경우, $P(X=a)$ 는 언제나 0이다. 따라서 pmf와 같은 개념은 없으나, 이와 비슷한 probability distribution function(PDF)를 이용하여 정의한다. 즉, 다음을 임의의 정의역 내 부분집합 B에서 만족하는 함수 f를 pdf라 한다.

$$P(X \in B) = \int_B f(x)dx \quad (11)$$

위와 비슷하게, probability cumulative distribution F는 다음과 같이 정의된다.

$$F(a) = P(X \in (-\infty, a]) = \int_{-\infty}^a f(x)dx \quad (12)$$

이 때, 위 두 식에서 $\frac{d}{dx}F(x) = f(x)$ 가 성립함을 알 수 있다.

- Uniform Random Variable

Uniform Random Variable

Uniform Random Variable은 pdf f 가 parameter (a,b) 에 대해서 다음과 같이 주어지는 확률변수를 뜻한다.

$$f(x) = \frac{1}{|a - b|}, x \in (a, b) \quad (13)$$

- 감마 확률변수

감마 확률변수

감마 확률변수는 pdf f 가 parameter $\lambda > 0, \alpha > 0$ 에 대해서 다음과 같이 주어지는 확률변수를 뜻한다.

$$f(x) = \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \quad (14)$$

이 때,

$$\Gamma(\alpha) = \int_0^{\infty} e^{-x} x^{\alpha-1} dx \quad (15)$$

이다. $\alpha = 1$ 인 경우, 이를 exponential random variable이라고 한다.

- 정규확률변수

정규확률변수

정규확률변수는 pdf f 가 parameter μ, σ 에 대해서 다음과 같이 주어지는 확률변수를 뜻한다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad (16)$$

- expectation

기대값(Expectation)

이산 확률변수 X 에 대해서 기대값 $E[X]$ 는 다음과 같다.

$$E[x] = \sum_{x=-\infty}^{\infty} xP(X = x) \quad (17)$$

비슷하게, 연속 확률변수 X 에 대해서 기대값 $E[X]$ 는 다음과 같다.

$$E[X] = \int_{-\infty}^{\infty} xP(X = x)dx \quad (18)$$

어떤 특정 함수 $g(X)$ 의 기대값은 x 대신 $g(x)$ 를 이용하여 계산하면 된다.

Statistical Values of Random Variables II

- variance

Variance

확률변수 X 에 대해서 Variance $\text{var}(X)$ 는 다음과 같다.

$$\text{Var}(X) = \sigma_X^2 = E[X^2] - E[X]^2 \quad (19)$$

- covariance/correlation

Covariance

확률변수 X, Y 에 대해서 Covariance $\text{Cov}(X, Y)$ 는 다음과 같다.

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \quad (20)$$

Correlation

확률변수 X, Y 에 대해서 Correlation $\text{Corr}(X, Y)$ 는 다음과 같다.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad (21)$$

Outline

- 1 확률의 정의
- 2 확률변수
- 3 조건부확률의 활용**
- 4 Different Perspectives on Probability

조건부확률을 이용한 기대값 계산 I

두 확률분포 X, Y 에 대해서 다음이 성립한다.

$$E[X] = E[E[X|Y]] \quad (22)$$

이산 확률변수의 경우와 연속 확률변수의 경우 각각 다음과 같다.

$$E[X] = \sum_y E[X|Y=y]P(Y=y) \quad (23)$$

$$E[X] = \int_y E[X|Y=y]P(Y=y)dy \quad (24)$$

이는 다음과 같이 증명할 수 있다.

조건부확률을 이용한 기대값 계산 II

$$\sum_y E[X|Y=y]P(Y=y) = \sum_y \sum_x xP(X=x|Y=y)P(Y=y) \quad (25)$$

$$= \sum_y \sum_x x \frac{P(X=x, Y=y)}{P(Y=y)} P(Y=y) \quad (26)$$

$$= \sum_y \sum_x xP(X=x, Y=y) \quad (27)$$

$$= \sum_x x \sum_y yP(X=x, Y=y) \quad (28)$$

$$= \sum_x xP(X=x) = E[X] \quad (29)$$

예시로, Randomized 퀵소트를 생각해 보자.

조건부확률을 이용한 분산 계산

모든 확률변수 X, Y 에 대해서 다음이 성립한다.

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]) \quad (30)$$

증명은 다음과 같다. 먼저, $E[\text{Var}(X|Y)]$ 는

$$E[\text{Var}(X|Y)] = E[E[X^2|Y] - E[X|Y]^2] \quad (31)$$

$$= E[E[X^2|Y]] - E[E[X|Y]^2] \quad (32)$$

또, $\text{Var}(E[X|Y])$ 는

$$\text{Var}(E[X|Y]) = E[E[X|Y]^2] - (E[E[X|Y]])^2 \quad (33)$$

$$= E[E[X|Y]^2] - (E[X])^2 \quad (34)$$

따라서 둘을 더하면

$$E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]) = E[X^2] - E[X]^2 = \text{Var}(X) \quad (35)$$

조건부확률을 이용한 확률 계산 I

어떤 사건 E 에 대해서, E 가 일어나면 1이고 아니면 0인 확률변수 X 를 생각하자. 이 때, $E[X] = P(E)$ 이고, 임의의 확률변수 Y 에 대해서 $E[X-Y=y] = P(E-Y=y)$ 이다. 이와 같은 formulation은 자명해 보이지만, 의외로 유용하게 쓰이는 경우가 있다. 예제를 통해서 살펴보겠다.
두 서로 독립인 확률변수 X, Y 에 대해서, 각각 pdf가 f_X, f_Y 라고 할 때 $P(X|Y)$ 를 생각해 보자. 이 때, Y 에 대한 조건부확률을 생각하여 다음과 같이 계산할 수 있다.

조건부확률을 이용한 확률 계산 II

$$P(X < Y) = \int_{-\infty}^{\infty} P(X < Y | Y = y) f_Y(y) dy \quad (36)$$

$$= \int_{-\infty}^{\infty} P(X < y | Y = y) f_Y(y) dy \quad (37)$$

$$= \int_{-\infty}^{\infty} P(X < y) f_Y(y) dy \quad (38)$$

$$= \int_{-\infty}^{\infty} F_X(y) f_Y(y) dy \quad (39)$$

두 번째 예시로, 베르누이 확률변수의 합의 분포에 대해서 생각해 보자. 즉, 서로 독립인 베르누이 확률변수 X_i 가 있다고 할 때, $\sum X_i$ 의 pmf를 생각해 보자. 이 때 이 pmf를 다음과 같이 쓸 수 있다.

$$P_k(j) = P\left(\sum_{i=1}^k X_i = j\right) \quad (40)$$

조건부확률을 이용한 확률 계산 III

다음의 식을 만족함을 보일 수 있다.

$$P_k(j) = pP_{k-1}(j-1) + (1-p)P_{k-1}(j) \quad (41)$$

Outline

- 1 확률의 정의
- 2 확률변수
- 3 조건부확률의 활용
- 4 Different Perspectives on Probability

Two Perspectives on Probability

확률의 해석에는 크게 두 사조가 있다.

- 빈도주의 (Frequentism)
- 베이지안 (Bayesianism)

Frequentism vs Bayesianism I



Frequentism vs Bayesianism I

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.



빈도주의자들은 확률을 **특정 사건이 일어날 빈도**로 파악한다. 빈도를 정의하기 위해서는 같은 환경에서의 실험을 반복할 수 있어야 하며, 이러한 반복을 무한히 반복하는 중 나오는 빈도의 극한으로 확률을 정의한다. 즉, 빈도주의자들에게 확률이란 **어떠한 고정된 값**이며, 동시에 데이터는 랜덤하게 뽑힌 값이다.

이와는 대조적으로, 베이지언에게는 확률은 어떠한 고정된 값이 아니라, 어떤 **명제의 확실한 정도를 나타내는 척도**이다. 따라서 확률은 데이터와 사전지식에 의거해서 지식을 수치화하는 하나의 체계이며, **고정된 값이 아니다**. 대조적으로, 데이터는 관측된 값으로 그것은 고정된 값이다.

MLE vs MAP I

두 접근의 차이를 알아보기 위해서, 어떤 별의 밝기를 측정하는 실험을 생각해 보겠다. n 번의 관측을 통해서 밝기 F_i 와 에러¹ e_i 를 안다고 생각하자. 즉,

$$P(D_i|F = F_t) = \frac{1}{\sqrt{2\pi e_i^2}} e^{\frac{-(F_i - F_t)^2}{2e_i^2}} \quad (42)$$

이다. 이 때, 관측된 D_i 들에 대해서 다음과 같이 likelihood 함수를 생각할 수 있다.

$$L(D|F = F_t) = \prod P(D_i|F = F_t) \quad (43)$$

이를 최소한으로 만드는 F_t 를 구하는 것을 **Maximum Likelihood Estimation**이라고 한다.

$$\frac{dL}{dF_t} = 0 \quad (44)$$

MLE vs MAP II

의 근 $F_{estimate}$ 를 구하고, 이것을 F_t 로 간주한다. 빈도주의자들에 따르자면 관측한 데이터가 충분히 많으면 이 진술은 참이 되며, 이를 대수의 법칙 (Law of Large Numbers)라 한다.

이것과 대조적으로, 베이지안은 $P(F = F_t|D)$ 가 최대가 되는 F_t 를 찾고자 한다. 관측은 우리에게 $P(D|F = F_t)$ 를 알게 해주므로, 원하는 F_t 를 구하기 위해서 베이즈의 법칙을 사용한다. 즉,

$$P(F = F_t|D) = \frac{P(D|F = F_t)P(F = F_t)}{P(D)} \quad (45)$$

이다. 여기서 각 항들의 의미를 살펴보면

- $P(F = F_t|D)$: Posterior 확률
- $P(D|F = F_t)$: Likelihood
- $P(F = F_t)$: Model prior / Prior 확률
- $P(D)$: data probability

¹정규분포에서의 표준편차라고 생각해야 한다. 오차-관측값 = 참값이 되는 오차가 아니다.